Single-cell RNA sequencing worksheet
Day8 part 1 (Project A)

Worksheet overview:

In this worksheet we will take the Seurat objects we integrated yesterday and begin visualizing the distribution of individual cells in UMAP plots and identify differentially expressed genes which is the first step towards annotating cell types.

Covered in this worksheet:

1) Clustering integrated data and generating UMAP plot of single cells
2) Identifying differentially expressed genes across clusters
3) Visualizing gene expression across UMAP clusters
4) Rename clusters by putative cell types

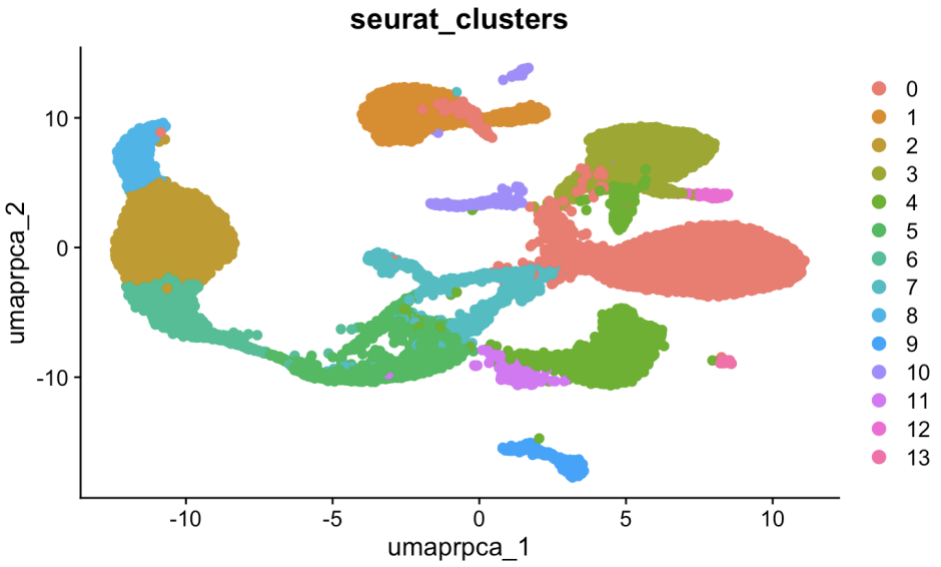
Code:

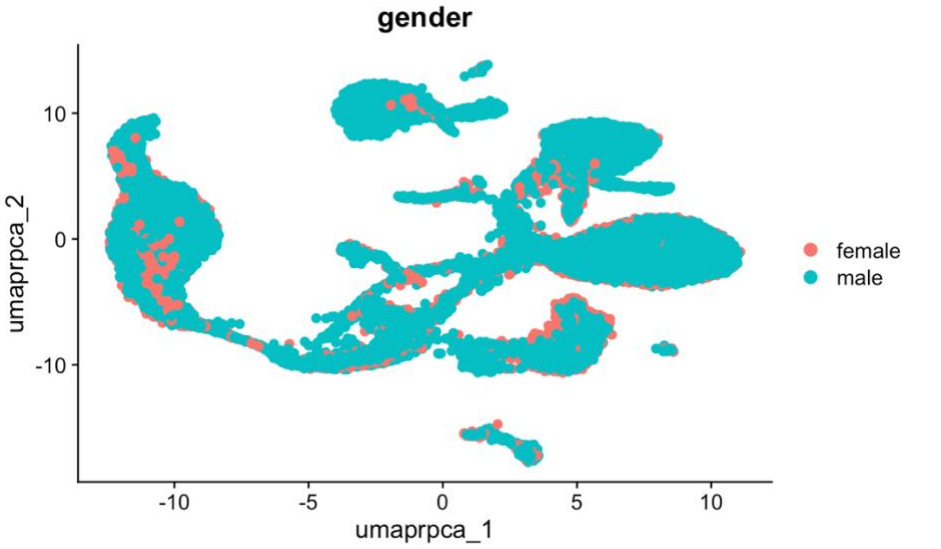
1) Clustering integrated data and generating UMAP plot of single cells. This is the primary step in visualizing single-cell data and is extremely subjective. The parameters in the RunUMAP() function (n.neighbors, min.dist, and dims) alter the distribution (clustering) of cells in this low-dimensional space. Often parameters are tuned to cluster cells according to what we know about the samples, the tissues the samples were derived from, and what cell types we expect to be present in our dataset. We want to achieve sufficient separation of clusters so that distinct cell types form distinct clusters, but we want to avoid "overfitting" our data such that the UMAP contains multiple distinct clusters of similar cell types. The FindClusters() function does nothing to change the visualization, but will change the number of actual clusters that Seurat infers are in the data.

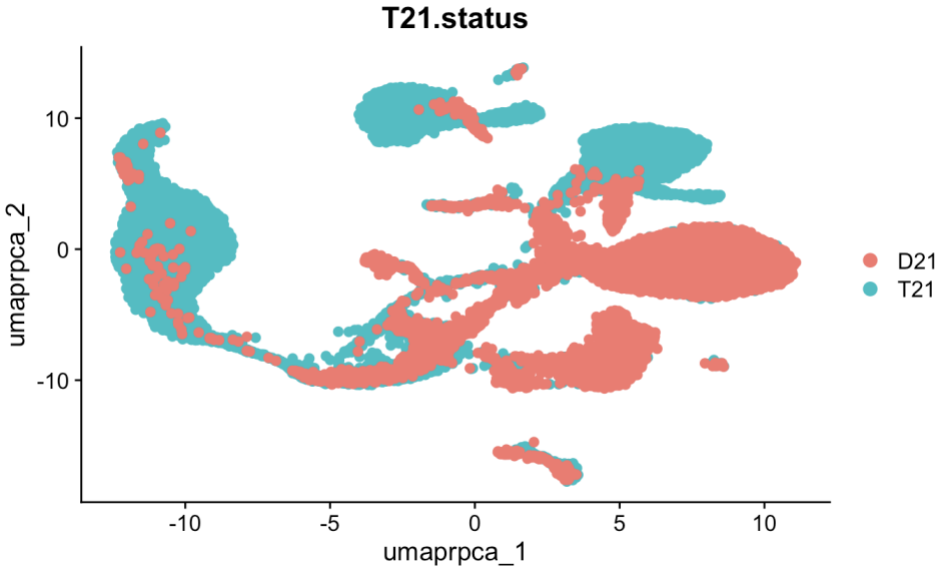
```
### Clustering data and generating UMAP
```{r}
obj <- FindNeighbors(obj, reduction = "integrated.rpca", dims = 1:30)
obj <- RunUMAP(obj, reduction = "integrated.rpca", dims = 1:30, reduction.name = "umap.rpca", n.neighbors = 100, min.dist = 0.5)
obj <- FindClusters(obj, resolution = 0.2, cluster.name = "rpca_clusters")
```

   a. Visualizing UMAP reduction. Plotting 3x identical UMAPs where the only change are the way the cells are colored – either by the Seurat-assigned clusters, or the gender/T21.status metadata designations we assigned earlier.

```
DimPlot(obj, pt.size = 2, group.by = 'seurat_clusters')
DimPlot(obj, pt.size = 2, group.by = 'gender')
DimPlot(obj, pt.size = 2, group.by = 'T21.status')
```
```


seurat_clusters
umaprpca_2
umaprpca_1
10
0
-10
-10
-5
0
5
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13


gender
umaprpca_2
umaprpca_1
10
0
-10
-10
-5
0
5
10
female
male


T21.status
umaprpca_2
umaprpca_1
10
0
-10
-10
-5
0
5
10
D21
T21

2) Identifying differentially expressed genes across clusters (which takes a long time) and between just cluster '1' and cluster '2'

```
# Differentially expressed genes
```{r}
# Identifying differentially expressed genes across all clusters
joint.layers.obj <- JoinLayers(obj)
all.clusters.diff.genes <- FindAllMarkers(joint.layers.obj, assay = "RNA")

# Identifying differentially expressed genes between cluster 1 and cluster 2
cluster1.cluster2.diff.genes <- FindMarkers(joint.layers.obj, ident.1 = '1', ident.2 = '2')
```

a. Exploring the table of differentially expressed genes

- Looking at the top 10 genes from each cluster based on average fold change

```
```{r}
all.clusters.diff.genes %>%
  group_by(cluster) %>%
  top_n(wt = avg_log2FC, n = 10)
```
```

A tibble: 390 × 7 Groups: cluster [39]

| p_val <dbl> | avg_log2FC <dbl> | pct.1 <dbl> | pct.2 <dbl> | p_val_adj <dbl> | cluster <fctr> | gene <chr> |
|---|---|---|---|---|---|---|
| 0.000000e+00 | 5.284415 | 0.282 | 0.016 | 0.000000e+00 | 0 | IL1RL1 |
| 2.613904e-75 | 4.646371 | 0.049 | 0.003 | 9.567149e-71 | 0 | PTHLH |
| 1.933341e-42 | 6.419573 | 0.017 | 0.000 | 7.076220e-38 | 0 | AC003684.1 |
| 5.493918e-38 | 4.661391 | 0.017 | 0.000 | 2.010829e-33 | 0 | TRBV10-1 |
| 9.300000e-33 | 5.096622 | 0.021 | 0.001 | 3.403893e-28 | 0 | TEPP |
| 5.853083e-25 | 4.967618 | 0.014 | 0.001 | 2.142287e-20 | 0 | AC108105.1 |
| 6.041299e-17 | 4.656507 | 0.015 | 0.002 | 2.211176e-12 | 0 | AC013460.1 |
| 2.614633e-16 | 4.671935 | 0.012 | 0.001 | 9.569817e-12 | 0 | AC105446.1 |
| 1.447907e-15 | 5.532648 | 0.012 | 0.001 | 5.299485e-11 | 0 | SNHG18 |
| 1.579633e-12 | 4.679966 | 0.011 | 0.001 | 5.781616e-08 | 0 | ITGA2 |

1–10 of 390 rows Previous 1 2 3 4 5 6 … 39 Next

- Looking at the top 10 genes from each cluster based on percent of cells expressing that gene in each cluster

```
```{r}
all.clusters.diff.genes %>%
  group_by(cluster) %>%
  top_n(wt = pct.1, n = 10)
```
```

A tibble: 629 × 7 Groups: cluster [39]

| p_val <dbl> | avg_log2FC <dbl> | pct.1 <dbl> | pct.2 <dbl> | p_val_adj <dbl> | cluster <fctr> | gene <chr> |
|---|---|---|---|---|---|---|
| 0.000000e+00 | 1.7895017 | 1.000 | 0.852 | 0.000000e+00 | 0 | RPL10 |
| 0.000000e+00 | 1.5686801 | 0.992 | 0.862 | 0.000000e+00 | 0 | RPL39 |
| 0.000000e+00 | 1.6283693 | 0.995 | 0.921 | 0.000000e+00 | 0 | RPL30 |
| 0.000000e+00 | 1.5325173 | 1.000 | 0.980 | 0.000000e+00 | 0 | RPL41 |
| 2.217691e-299 | 1.4639538 | 0.992 | 0.902 | 8.116973e-295 | 0 | RPL11 |
| 2.808537e-262 | 1.3225002 | 0.992 | 0.903 | 1.027953e-257 | 0 | RPL32 |
| 2.807130e-261 | 1.1920336 | 0.999 | 0.939 | 1.027438e-256 | 0 | RPLP1 |
| 1.834952e-256 | 1.3483953 | 1.000 | 0.820 | 6.716107e-252 | 0 | B2M |
| 5.002497e-249 | 1.3027374 | 0.994 | 0.940 | 1.830964e-244 | 0 | RPS12 |
| 2.535600e-144 | 1.0825314 | 0.993 | 0.988 | 9.280549e-140 | 0 | FTL |

1–10 of 629 rows Previous 1 2 3 4 5 6 … 63 Next

- Can also identify differentially expressed genes by metadata

```
differentially expressed genes by metadata
```{r}
joint.layers.obj <- SetIdent(joint.layers.obj, value =  'T21.status')
diff.genes.tmp1 <- FindAllMarkers(joint.layers.obj)
```
```
```

- Creating a subset of the object to only contain cells in cluster 1 and then identifying differentially expressed genes by metadata

```
joint.layers.obj <- SetIdent(joint.layers.obj, value =  'seurat_clusters')
joint.layers.obj.subset <- subset(joint.layers.obj, idents = '1')
joint.layers.obj.subset <- SetIdent(joint.layers.obj.subset, value =  'T21.status')
diff.genes.tmp1 <- FindAllMarkers(joint.layers.obj.subset)
```

a. Genes differentially expressed between cluster '1' and cluster '2'

3) Visualizing gene expression on UMAP plot

```
```{r}
DefaultAssay(joint.layers.obj) <- "RNA"

FeaturePlot(joint.layers.obj, features = 'SLC4A1')
FeaturePlot(joint.layers.obj, features = 'FCGR3B')
```
```